Replicating Multi-Year Polling Trends: Comparing 1992 and 2020 Presidential Elections

Completed for the Certificate in Scientific Computation and Data Sciences
Spring 2023

Matthew Dauber
Bachelor's Degree
Government
College of Liberal Arts

Shannon Bow O'Brien, Ph.D.
Associate Professor of Instruction
Government

Replicating Multi-Year Polling Trends: Comparing 1992 and 2020 Presidential Elections

Abstract

The purpose of this project was to attempt to replicate race and population density voting trends from the 1992 and 2020 presidential elections using polling conducted by *ABC News* and *The Washington Post*, collected from the Roper Center, for both years. The author cleaned and processed 12 datasets, merging them into a single dataset. Much of this data was legacy data, from defunct files that had to be modified to be forward compatible with modern data. The author used this data to run logistic regression models using R Studio IDE and the dplyr and rio packages. This model defined voting for a Democratic candidate as the success case, and used White voters from non-densely populated regions as the intercept. The findings from the polling data suggested the Democratic candidates in both election years won the Black vote within metropolitan and suburban areas. This replicated similar results from other pre-existing research on these presidential races.

Introduction

Presidential election polls serve as a way to get in touch with the political thoughts and feelings of America. When discussing the 1992 and 2020 elections, people assume certain trends to be accurate. However, making claims based on those trends without confirming the pattern of the trend in question can result in misleading conclusions. In 1992 and 2020, exit poll data showed that the Democratic campaign carried a significant amount of the Black and the metropolitan vote. This paper will use polling data from *ABC News* and *The Washington Post* leading up to both elections to replicate this effect. After conducting a logistic regression analysis, this paper suggests that the Black vote had a significant effect on the results of the election while the metropolitan vote has a smaller but still significant impact.

Presidential elections are a critical juncture of American democracy. The president is one of the three checks and balances that we are taught as children. Each president has a great impact that extends past their own term, into the future. The presidential election, as the decider of such an important role, garners immense focus and attention worldwide. Given the level of significance attached to the presidency, it is only suitable that that same level of focus should be directed towards examining/investigating the path taken to assuming the mantle of the presidency, with one example of this being presidential election polling.

Polling data helps voters make informed choices. Humans like to win - winning produces serotonin, winning feels good, being on the winning side feels good - even if winning in this case is just when a favored candidate is polling well. Polling data gives us metrics through which we can assess our options. These metrics are valuable in multiple ways, even beyond purely academic or statistical interests.

First, this information helps supporters feel validated in their voting choices. Voters want to know how their favorite candidate is doing, how likely they are to win candidacy. They want to know if which party is winning or losing, at national, state, and local levels. Polling data can also, at times, help voters overcome the paradox of voting. By helping individuals to feel that their support has a direct impact, a potential voter who otherwise wouldn't have voted may decide to vote upon realizing a valued race is close.

Since polls have become so important, supply has arisen to meet this demand. Over the last century, news organizations have filled this niche, conducting polls and including that information alongside other offerings. Additionally, companies value having the most accurate and robust polling data for the sake of advancing their own capitalist interests.

The other groups interested in election polling data are the candidates themselves, and their campaigns. Polling data helps candidates and their campaigns target voters or see if their message resonates. Polling data can also direct advertising and spending efforts, both on when the campaign needs to take action, and what shape those actions should take.

Polling data has become critical to a vast number of people, and scientific polling has grown concurrently into a field of study and academic area to meet this need. It is collected over the course of a campaign, with national datasets using random sampling to create snapshots of attitudinal beliefs amongst the nation's population when done correctly. However, given the potential problems inaccurate polls would present, it is important to validate polling data after an election has concluded, both to ensure that the polls accurately described the trends of each election, and that the trends we ascribe to that election are upheld by polling data.

Polling data helps voters to better determine the more popular candidates, as well as major issues for any given survey period. It is often collected at different points during a political campaign as a way to gauge public attitudes. National datasets rely upon random sampling to extrapolate the broader population. When correctly collected, they will accurately gauge attitudinal beliefs from their sample. Many people look at it as a way to gauge a candidate's popularity, viability, as well as the likelihood of success. Scientific polling has developed as a field within the last century. It is a dynamic academic area that utilizes many methods and techniques. However, unusual results can emerge when undersampled groups are accidentally overrepresented within populations.

The University of Southern California Dornsife/Los Angeles Times Daybreak Poll(U.S.C./LAT Poll) is one such example of a case where upon comparison between this poll and other polls, including exit polling data, it became clear that something was off. The U.S.C./LAT Poll made two choices that on their own, were not particularly problematic, but when combined, led to one Black Trump supporter being "weighted as much as 30 times more than the average respondent, and as much as 300 times more than the least-weighted respondent”(Cohn, 2016). The first of these choices was asking how people voted in the 2012 election, and weighting based on that past vote. The second problematic decision was weighting with smaller groupings than other comparable polls. Instead of weighting to be representative of the broader population for groups like the “number of men, or or the right number of people 18 to 29,” the U.S.C./LAT Poll weighted for much smaller and more specific categories such as “18-to-21-year-old men”(Cohn, 2016). This level of specificity meant that more weighting would be required, which led to extravagant weighting being given “to particularly underrepresented voters — like 18-to-21-year-old black men”(Cohn, 2016).

Despite the U.S.C./LAT Poll having “emerged as the biggest polling outlier of the [2016] presidential campaign”, the source of the discrepancy was only discovered due to the poll’s transparency and publication of their data set and documentation, allowed for the replication of the analysis and let others test the data with different weights(Cohn, 2016). This is but one example of how important it is to double check the results and conclusions we draw from polls.

This project examines the Democratic support by race and population density in both the 1992 and 2020 presidential elections. Both elections are ones where an incumbent president lost to a challenger and where authors such as Wickham (2002), Hunnicutt et al. (2020), Judd and Hinze (2019) have asserted that Black turnout played a critical role in these results. Additionally, Schneider (1992), Judd and Hinze (2019), and Boak and Fingerhut (2020) impress the importance of urban and suburban areas in deciding these elections.

Clinton was very popular with Black voters and “his ability to capture 82 percent of the African American vote was crucial to his victory” (Judd and Hinze 2019, 250). Clinton’s popularity lasted beyond the leadup to his election. Wickham describes a poll done in 1998 by a Black owned survey and research group “found that Clinton’s approval rating among African Americans (93%) was higher than that of the Reverend Jesse Jackson (89%)”(Wickham, 2002). Given Clinton’s extreme popularity among African Americans, “ the Clinton campaign decided to concentrate on appealing to the white suburban middle class,” taking the urban inner cities for granted, believing they “would support him anyways because they had no place else to go” (Judd and Hinze 2019, 250). This was especially effective given that “in 1988, the suburbs accounted for 48 percent of the vote”(Schneider, 1992). Additionally, Judd and Hinze write that the suburbs “tended to turn out for elections at a relatively high rate” (Judd and Hinze, 2019). This situation cemented the suburbs as a pivotal part of the election and a primary focus of Clinton’s campaign’s attention. In the end, “Clinton succeeded by winning back many of the white suburban voters who had deserted the party in 1980” (Judd and Hinze, 2019).

Biden had a similar popularity with Black voters during the 2020 election, where he “received the support of over 92% of Black voters” (Igielnik et al., 2021). In order to achieve a higher turnout in major metropolitan areas, “in the final weeks, the Biden campaign turned to get-out-the-vote events.”(Hunnicutt et al., 2020) The Biden campaign’s “better job of driving turnout of Black voters, the Democratic Party’s most loyal constituency,” led to “election officials in Detroit said turnout was the highest in 20 or 30 years” and higher turnout “over 2016 in Milwaukee and Philadelphia”(Hunnicutt et al., 2020). Biden further appealed to Black voters, convincing them “that their concerns would be heard in his administration,” speaking “often of his close relationship with Obama” and selecting “U.S. Senator Kamala Harris as his running mate, the first Black woman to join a major-party ticket”(Hunnicutt et al., 2020). Boak and Fingerhut emphasize the importance of densely populated areas, stating that “he [Biden] outpaced Trump in the suburbs, 54% to 44%, and dominated with roughly two-thirds of voters in urban areas”(Boak and Fingerhut, 2021) The Pew Research Center also describes how “the political split between America’s rural areas and its suburban and urban locales remained substantial in 2020,” with Biden garnering 54% of the suburban votes, and Trump receiving 65% of the rural voters (Igielnik et al., 2021).

Materials and Methods

Table 1. The Twelve Samples - Sources and Data Collection Dates

No.	Title	Collection Dates	Sample Size
1	ABC News/Washington Post Poll: 1992 Presidential Election	July 10 - 14, 1992	1001
2	ABC News/Washington Post Poll: Election Tracking Poll II	August 7 - 11, 1992	1009
3	ABC News/Washington Post Poll: Election Tracking Poll III	August 12 - 16, 1992	1003
4	ABC News/Washington Post Poll: Election Tracking Poll III	August 14 - 18, 1992	1003
5	ABC News/Washington Post Poll: Election Tracking Poll VI	September 9 - 13, 1992	1002
6	ABC News/Washington Post Poll: Election Tracking Poll VII	September 16 - 20, 1992	1003
7	ABC News/Washington Post Poll: Trump/2020 Democratic Presidential Nomination/Iran Drone Strike	January 20 - 23, 2020	1004
8	ABC News/Washington Post Poll: Trump/2020 Democratic Primary Candidates	February 14 - 17, 2020	1066
9	ABC News/Washington Post Poll: March 2020 Coronavirus	March 22 - 25, 2020	1003
10	ABC News/Washington Post Poll: Trump/Coronavirus/2020 Presidential Election	May 25 - 28, 2020	1001
11	ABC News/Washington Post Poll: July 2020 Coronavirus	July 12 - 15, 2020	1006
12	ABC News/Washington Post Poll: August 2020 Coronavirus	August 12 - 15, 2020	1001

The purpose of this paper is to test the credibility of assumed trends regarding race and population density from 1992 and 2020 using representative same organization national election polling data. The author chose to narrow the data selection down by examining the available polling data from both years, and selecting based on organization. In order to reduce the scope of this project to better match the timeline, this author chose to access only polling data available through the Roper Center. During this process, this paper originally compared polling data from both years based on similarities between the elections, most notably because both elections had incumbent Republican candidates who were defeated. As such, keeping the survey data to a single originating organization, with as similar a question set as possible, was a priority. The survey selection process hinged on whether the poll directly addressed the question of presidential vote, and the availability of the data.

While the datafiles for the 1992 polling data were available, the majority was only available in American Standard Code for Information Interchange, or ASCII. The data had all been deposited, but since the codebooks had not been included in the data deposit, there was no way to decipher the data and identify the variables. The only polling datasets with codebooks were the datasets from *ABC News* and *The Washington Post*. Deciphering the data from *ABC News* and *The Washington Post* was further aided by knowledge of the question sets. It was fortunate that both *ABC News* and *The Washington Post* had polling data for both years with similar if not identical questions, and with codebooks attached to ensure reliability of the questions. These datasets were available in multiple file types, including comma separated value (CSV) files, and file formats compatible with SPSS statistical software. The selection process resulted in twelve surveys from *ABC News* and *The Washington Post*, with six from 1992 and six from 2020, converted from SPSS to CSV in section one of the appendix using the rio library.

The original research strategy was to include all of the available demographic information as explanatory variables, accompanying presidential vote, the response variable. Despite this lofty goal, this author eventually winnowed down the demographic variable list to just population density and race, which occurs in section five of the appendix. Relating these two variables to trends from each election was the most achievable within the timeline of this project. For the 1992 survey data, the presidential vote variable was originally coded categorically, with 1 for George Bush, 2 for Bill Clinton, 3 for Ross Perot, 0 for other candidate, while 7, 8, and 9 indicated a refusal to vote, no opinion, and refusing the question, respectively. Since this paper focuses on voting trends between the two primary parties, the author chose to group Ross Perot responses alongside responses of other candidate, refusal to vote, no opinion, and question refusal when recoding responses as NAs, which occurs in section four of the appendix. In preparing the data for the logistic regression, this author recoded George Bush as 0, and Bill Clinton as 1, defining a vote for Clinton as the success case in section 8 of the appendix.

For the 2020 survey data, the presidential vote variable was measured and coded similarly to the 1992 data, sans a category naming a specific independent candidate, and the non-primary candidate responses being numbered 3, 4, and 5 instead. This author recoded Donald Trump as 0, and Joe Biden as 1, defining a vote for Biden as the success case, also in section 8 of the appendix. The other responses were recoded as NAs. The race variable for 2020 was also categorical, with options of Hispanic, White, Black, White Hispanic, Black Hispanic, Asian, other, along with the accompanying no opinion and refused. Due to the structure of the poll, the question was split into two columns, and the author chose to use the one that delineated only between Hispanic, White, Black, and Asian, leaving out the specification between Black and White Hispanic people. This was done in order to make 1992 and 2020 more comparable, as per the author’s original thesis, since the 1992 race variable only delineates between Hispanic, White, and Black. The population density variable, metro, also changed between 1992 and 2020. The 1992 polls categorized respondents as metropolitan or non-metropolitan, based on their zip code, while the 2020 polls categorized respondents as urban, suburban, or rural. In order to make these directly comparable, the author chose to merge the urban and suburban response codes into metropolitan, and recoded rural into non-metropolitan. The dplyr package adds a “grammar of data manipulation, providing a consistent set of verbs” to help “solve the most common data manipulation challenges” (Wickham et al., n.d.) These “verbs” or methods, allow the data scientist to use the pipe operator, %>%, to chain together multiple individual modifications and transformations on a data frame. Some of the functions contained in the dplyr package are mutate, which “adds new variables that are functions of existing variables,” select, which “picks variables based on their names,” and filter, which “picks cases based on their values” (Wickham et al., n.d.). The dplyr package proved essential to the preparation, cleaning, and processing of this dataset.

Based on the categorical nature of the data, the author decided to make use of the logistic regression model for this statistical analysis, conducted through the R Studio IDE. The logistic regression model, or ‘logit’, was chosen because the response variable, presidential vote, is a binomial categorical variable. The combined dataset was split by year in order to run two separate models, one on the 1992 data, and one on the 2020 data. The success case for both years was defined as a vote for the Democratic candidate in section 8 of the appendix. Since a logit is “the logarithm of the odds”, the logistic regression model will describe the odds of the success case occurring compared to the intercept case (“4.1.1 The Logistic Regression Model,” 2018). In the models constructed in this paper, the intercept case is a White voter from a non-densely populated region. This then serves to contrast the explanatory variables, race and population density.

Results & Discussion

Table 2. The Effect of Race and Population Density on Presidential Vote

	1992 Election Polling Data	2020 Election Polling Data
Intercept	0.12(0.07)	-0.85***(0.84)
Race: Black	2.13***(0.19)	2.83***(0.25)
Race: Hispanic	0.82***(0.22)	0.66***(0.10)
Race: Asian	-	0.64*(0.25)
Metro: Metropolitan	0.16*(0.08)	0.83***(0.09)
n	3459	3999

Note: Entries are logistic regression coefficients, with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Does the 1992 and 2020 polling data support the exit polling trend from those years of Black voters and voters from densely populated areas supporting Democratic candidates?

This paper's findings show that our perception of trends during the 1992 and 2020 elections with regard to the Black vote and vote of those from densely populated areas is supported by polling data from both years. More specifically, for both 1992 and 2020, there is significance between Black voters and voting for Democratic candidates, and between voters from densely populated regions and voting for Democratic candidates.

Table 2 reports the effects of race and population density on presidential vote. The findings in Table 2, Model 1992 shows that race strongly correlates with presidential vote ($\beta_{intercept1992} = 0.12$, $p < 0.1$, $\beta_{black1992} = 2.13$, $p < 0.0001$, $\beta_{hispanic1992} = 0.82$, $p < 0.001$). A Black voter in the 1992 election had 8.4 times the odds of a White voter from a non-metropolitan area of having voted for a Democratic candidate, while a Hispanic voter had 2.3 times the odds of a White voter from a non-metropolitan area of having voted for a Democratic candidate in the 1992 election. Additionally, originating from a densely populated area correlates with presidential vote ($\beta_{metropolitan} = 0.16$, $p < 0.01$). This logistic regression coefficient gives a voter from a densely populated area 1.17 times the odds of voting for a Democratic candidate as a White voter from a non-densely populated area in 1992. These findings support Judd and Hinze's statistic that Clinton captured "82 percent of the African American vote" (Judd and Hinze 2019). Clinton's focus on appealing to densely populated areas is evident, as was his success at "winning back many of the White suburban voters who had deserted the [Democratic] party in 1980", but population density level has less of an impact on having voted for the Democratic candidate than being a Hispanic voter (Judd and Hinze 2019).

The findings in Table 2, Model 2020 show that race strongly correlates with presidential vote($\beta_{intercept2020} = -0.85$, $p < 0.0001$, $\beta_{black2020} = 2.83$, $p < 0.0001$, $\beta_{hispanic2020} = 0.66$, $p < 0.0001$, $\beta_{asian2020} = 0.64$, $p < 0.01$). A Black voter in the 2020 election had 17.02 times the odds of a White voter from a non-densely populated area of having voted for a Democratic candidate. For the same election, a Hispanic voter had 1.93 times the odds of a White voter from a non-densely populated area of having voted for a Democratic candidate in the 2020 election, while an Asian voter had 1.89 times the odds of a White voter from a non-densely populated area of having voted for a Democratic candidate in this election. These findings indicate that the effect of being Hispanic on presidential vote is comparable to the effect of being Asian on presidential vote for the 2020 election. Additionally, during this election, voters of any race from densely populated areas($\beta_{metropolitan2020} = 0.83$, $p < 0.0001$) had 2.28 times the odds of a White voter from a non-densely populated area of having voted for a Democratic candidate in the 2020 election. As in 1992, this model's findings remain consistent with the election results. The high odds of a Black voter voting for Biden in 2020 are supported by him receiving "the support of over 92% of Black voters" (Igielnik et al., 2021). The higher likelihood of a voter from a densely populated region voting for Biden is consistent with how "he outpaced Trump in the suburbs, 54% to 44%, and dominated with roughly two-thirds of voters in urban areas" (Boak and Fingerhut, 2021) .

Year by year, surveys and polls have developed as a field, growing to meet the demands of voters, parties, candidates, campaigns, and corporations.The Business Research Company states that "the global public opinion and election polling market size will grow from $7.94 billion in 2022 to $8.15 billion in 2023" (*Public Opinion and Election Polling Market Size, Trends and Global Forecast to 2032*, n.d.). Polls use a multitude of techniques and methods to correctly gauge attitudinal beliefs about their subject population. Polling is prevalent in our lives/Polling has cemented itself as part of our lives. Validating the results of polls and surveys is one part of ensuring the accuracy of polling claims.

This paper's findings suggest that trends during the 1992 and 2020 elections with regard to the Black vote and vote of those from densely populated areas are supported by polling data from both years. Black voters had the strongest relationship with Democratic candidates, with Black voters having 8.4 and 17.02 times the odds of voting for a Democratic candidate as compared to a White voter from a non-densely populated region for the 1992 and 2020 elections, respectively. These odds ratios support the claims made by authors Judd and Hinze (2019) and Igielnik et al.. (2021) about the Black vote in both years. Voters from densely populated regions had a smaller but still significant relationship with Democratic candidates in both years. Voters of any race from densely populated areas had 1.17 times the odds in 1992, and 2.28 times the odds in 2020, of having voted for a Democratic candidate as compared to a White voter from a non-densely populated region. These odds ratios substantiate the claims made by Schneider (1992), Judd and Hinze (2019), and Boak and Fingerhut (2020). These results suggest that our perception of trends from 1992 and 2020 is indeed real, reinforcing the accuracy and reliability of polling data.

This project attempts to highlight the importance of examining and reexamining historical data to see how trends emerge or continue over time. There were a number of limitations within this current project that could be improved going forward. One of the primary improvements would be adding more data. Much of the 1992 data was inaccessible due to the missing codebooks for the ASCII data. Without these codebooks, it became impossible to utilize the data without running the risk of creating ecological fallacies. If those codebooks could be recovered, or the data deciphered independently, the additional data would make the logistic regression model more robust, and allow for additional confidence in this paper's results. Additionally, expanding to include data from multiple years would also allow for an examination of wider spanning trends. This would also require the identification of trends for each year, or selecting a specific overarching trend to examine, but could make for an interesting study. Another potential future research topic would be expanding the demographics considered in the logistic regression model and in the analysis. This would also require the identification of additional trends, but would allow the project to more fully describe the voting patterns of the included years.

Acknowledgements

This author would like to extend a huge thank you to Professor Shannon Bow O'Brien; without you none of this would have been possible.
Additionally, the author greatly appreciates the aid of graduate student Kyle Shen for answering statistics questions, and Megan Cockett for help editing.

Bibliography

4.1.1 The Logistic Regression Model. (2018). In A. Agresti, *An Introduction to Categorical Data Analysis* (p. 90). John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/utxa/detail.action?docID=5592838

ABC News, & Washington Post. (2010a). *ABC News/Washington Post Poll: 1992 Presidential Election*. https://doi.org/10.25940/ROPER-31086773

ABC News, & Washington Post. (2010b). *ABC News/Washington Post Poll: Election Tracking Poll II*. https://doi.org/10.25940/ROPER-31086774

ABC News, & Washington Post. (2010c). *ABC News/Washington Post Poll: Election Tracking Poll III*. https://doi.org/10.25940/ROPER-31086766

ABC News, & Washington Post. (2010d). *ABC News/Washington Post Poll: Election Tracking Poll III*. https://doi.org/10.25940/ROPER-31086775

ABC News, & Washington Post. (2010e). *ABC News/Washington Post Poll: Election Tracking Poll VI*. https://doi.org/10.25940/ROPER-31086769

ABC News, & Washington Post. (2010f). *ABC News/Washington Post Poll: Election Tracking Poll VII*. https://doi.org/10.25940/ROPER-31086770

ABC News, & Washington Post. (2020a). *ABC News/Washington Post Poll: August 2020 Coronavirus*. https://doi.org/10.25940/ROPER-31117650

ABC News, & Washington Post. (2020b). *ABC News/Washington Post Poll: March 2020 Coronavirus*. https://doi.org/10.25940/ROPER-31117260

ABC News, & Washington Post. (2020c). *ABC News/Washington Post Poll: Trump/2020 Democratic Presidential Nomination/Iran Drone Strike*. https://doi.org/10.25940/ROPER-31117054

ABC News, & Washington Post. (2020d). *ABC News/Washington Post Poll: Trump/Coronavirus/2020 Presidential Election*. https://doi.org/10.25940/ROPER-31117423

ABC News, & Washington Post. (2022a). *ABC News/Washington Post Poll: July 2020 Coronavirus*. https://doi.org/10.25940/ROPER-31117566

ABC News, & Washington Post. (2022b). *ABC News/Washington Post Poll: Trump/2020 Democratic Primary Candidates*. https://doi.org/10.25940/ROPER-31117150

Boak, J., & Fingerhut, H. (2021, April 20). AP VoteCast: How did Biden do it? Wide coalition powered win. AP NEWS. https://apnews.com/article/how-did-joe-biden-win-election-a493c68b6b947c5f90f36efef76d13c2

Cohn, N. (2016, October 12). How One 19-Year-Old Illinois Man Is Distorting National Polling Averages. *The New York Times*. https://www.nytimes.com/2016/10/13/upshot/how-one-19-year-old-illinois-man-is-distorting-national-polling-averages.html

Hunnicutt, T., Oliphant, J., Ax, J., & Renshaw, J. (2020, November 7). Biden's winning strategy: Flip Rust Belt Trump states and hold on tight. *Reuters*. https://www.reuters.com/article/us-usa-election-biden-insight-idUSKBN27N0OC

Igielnik, R., Keeter, S., & Hartig, H. (2021, June 30). *Behind Biden's 2020 Victory | Pew Research Center*. https://www.pewresearch.org/politics/2021/06/30/behind-bidens-2020-victory/

LANGER, G. (2020, September 15). *ABC News' Polling Methodology and Standards*. ABC News. https://abcnews.go.com/US/PollVault/abc-news-polling-methodology-standards/story?id=145373

Schneider, W. (1992, July). *The Suburban Century Begins—92.07*. https://www.theatlantic.com/past/docs/politics/ecbig/schnsub.htm

Judd, D. R., & Hinze, A. M. (2018). City politics: The Political Economy of Urban America (10th ed.). Routledge. https://doi.org/10.4324/9781315166018

Wickham, D. (2002). Bill Clinton and Black America. Random House Publishing Group.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (n.d.). Dplyr. Retrieved April 19, 2023, from https://dplyr.tidyverse.org/

Public opinion and election polling market size, trends and global forecast to 2032. (n.d.). Public Opinion And Election Polling Global Market Report 2023; The Business Research Company. Retrieved April 19, 2023, from https://www.thebusinessresearchcompany.com/report/public-opinion-and-election-polling-global-market-report

Appendix

R code

```
library(interactions)
library(sandwich)
library(ggplot2)
library(dslabs)
library(dplyr)
library(tidyverse)
library(stringr)
library(rio)

# section 1: loading data ----------------------------------------------------------
#creates a function that takes a folder as input, and converts all spss files into csv files.
#single use code
RIO_SPSS2CSV <- function(filepath) {
  setwd("C:\\Users\\mndau\\Documents\\R\\ElectionRPData\\") #this is the root dir where SPSS
data files/folders are located; .csv files will be stored in the same dir
  files <- list.files(path = filepath,
                      pattern = '.por',
                      recursive = TRUE) #recursive option to check all folders inside the root dir
  for (f in files) {
    convert(f,
            paste0(strsplit(f,
                            split = '.',
                            fixed = TRUE)[[1]][1],'.csv'))
  }
}

RIO_SPSS2CSV("C:\\Users\\mndau\\Documents\\R\\ElectionRPData\\")

#loading 12 datasets
```

```
filepath <- "C:\\Users\\mndau\\Documents\\R\\ElectionRPData\\"
files <- list.files(path = filepath,
                    pattern = '.csv',
                    recursive = FALSE) #recursive option to check all folders inside the root dir
setwd("C:\\Users\\mndau\\Documents\\R\\ElectionRPData\\")
d32 <- read.csv(files[1])
d33 <- read.csv(files[2])
d34 <- read.csv(files[3])
d15 <- read.csv(files[4])
d3 <- read.csv(files[5])
d2 <- read.csv(files[6])
d35 <- read.csv(files[7])
d13 <- read.csv(files[8])
d6 <- read.csv(files[9])
d11 <- read.csv(files[10])
d30 <- read.csv(files[11])
d31 <- read.csv(files[12])


# section 2: selecting potential variables prejoining -------------------------------
#selecting the desired columns/vars for each datasets prejoining
d2prep <- d2 %>% select(respo, project, date8,
                        partlean, q13_1, q13_1net,
                        q910, q909, q909a,
                        edubreak, colleduc, educnew,
                        income, income2, q918,
                        racenet, hisprace, q924,
                        q924net, abcnum, stcode,
                        stcode2, weight, pidwgt,
                        reg4, nreg4, censdiv,
                        ncensdiv, usr, nusr)  %>%
  mutate(q910=as.numeric(q910))

d3prep <- d3 %>% select(respo, project, date8,
                        partlean, q19_1, q19_1net,
                        q910, q909, q909a,
                        edubreak, colleduc, educnew,
                        income, income2, q918,
                        racenet, hisprace, q924,
                        q924net, abcnum, stcode,
                        stcode2, pidwgt, reg4,
                        nreg4, censdiv, ncensdiv,
                        usr, nusr) %>%
  mutate(q910=as.numeric(q910))

d6prep <- d6 %>% select(respo, date8, partlean,
                        q8, q8net, q910,
                        q909, q909a, edubreak,
                        colleduc, educnew, income,
                        income2, q918, racenet,
                        hisprace, q924, q924net,
                        abcnum, stcode, stcode2,
                        weight, reg4, nreg4,
                        ncensdiv, usr, nusr) %>%
  mutate(q910=as.numeric(q910))

d11prep <- d11 %>% select(respo, project, date8,
                          partlean, q8, q8net,
                          q910, q909, q909a,
                          edubreak, colleduc, educnew,
                          income, income2, q918,
                          racenet, hisprace, q924,
                          q924net, abcnum, stcode,
                          stcode2, weight, reg4,
                          nreg4,censdiv, ncensdiv,
                          usr, nusr) %>%
  mutate(q910=as.numeric(q910))

d13prep <- d13 %>% select(respo, project, date8,
                          partlean, q2, q2net,
                          q910, q909, q909a,
                          edubreak, colleduc, educnew,
                          income, income2, q918,
                          racenet, hisprace, q924,
                          q924net, abcnum, stcode,
                          stcode2, weight, reg4,
                          nreg4, censdiv, ncensdiv,
                          usr, nusr) %>%
  mutate(q910=as.numeric(q910))

d15prep <- d15 %>% select(respo, project, date8,
                          partlean, q2, q2net,
                          q910, q909, q909a,
                          edubreak, colleduc,
                          educnew,income, income2,
                          q918, racenet, hisprace,
                          q924, q924net, abcnum,
                          stcode, stcode2, weight,
                          reg4, nreg4, censdiv,
                          ncensdiv, usr, nusr) %>%
  mutate(q910=as.numeric(q910))
```

```
d35prep <- d35 %>% select(ID, WEEK, WP3, Z7,
              Z8, Z9, Z10, Z10A,
              Z11, Z11A, SEX, METRO,
              REGION, HWEIGHT, PWEIGHT, STATE)



d34prep <- d34 %>% select(ID, WEEK, AW4, Z7,
              Z8, Z9, Z10, Z10A,
              Z11, Z11A, SEX, METRO,
              REGION, HWEIGHT, WEIGHT, STATE)

d33prep <- d33 %>% select(ID, WEEK, AW4, Z7,
              Z8, Z9, Z10, Z10A,
              Z11, Z11A, SEX, METRO,
              REGION, HWEIGHT, PWEIGHT, STATE)

d32prep <- d32 %>% select(ID, WEEK, AW4, Z7,
              Z8, Z9, Z10, Z10A,
              Z11, Z11A, SEX, METRO,
              REGION, HWEIGHT, WEIGHT, STATE)

d31prep <- d31 %>% select(ID, WEEK, AW4, Z7,
              Z8, Z9, Z10, Z10A,
              Z11, Z11A, SEX, METRO,
              REGION, HWEIGHT, PWEIGHT, STATE)

d30prep <- d30 %>% select(ID, WEEK, AW4, Z7,
              Z8, Z9, Z10, Z10A,
              Z11, Z11A, SEX, METRO,
              REGION, HWEIGHT, PWEIGHT, STATE)


# section 3: creating joinlists and merging into 2 datasets ---------------------------------
#joining the prepped datasets together. first doing a batch of 2020, then a 1992 batch, then merging

joinlist2020 <- c("respo", "date8", "partlean",
          'q910', 'q909', 'q909a',
          'edubreak', 'colleduc', 'educnew',
          'income', 'income2', 'q918',
          'racenet', 'hisprace', 'q924',
          'q924net', 'abcnum', 'stcode',
          'stcode2', 'reg4', 'nreg4',
          'ncensdiv', 'usr', 'nusr')
```

```
#manual: project, weight, pidwgt, censdiv
d2020 <- full_join(d2prep,
                   d3prep,
                   by=c(joinlist2020,
                        "project", "censdiv",
                        "pidwgt", "q13_1"="q19_1",
                        "q13_1net"="q19_1net")) %>%
  full_join(d6prep,
            by=c(joinlist2020,
                 "weight", "q13_1"="q8",
                 "q13_1net"="q8net")) %>%
  full_join(d11prep,
            by=c(joinlist2020,
                 "project", "censdiv", "weight",
                 "q13_1"="q8", "q13_1net"="q8net")) %>%
  full_join(d13prep,
            by=c(joinlist2020,
                 "project", "censdiv", "weight",
                 "q13_1"="q2", "q13_1net"="q2net")) %>%
  full_join(d15prep,
            by=c(joinlist2020,
                 "project", "censdiv", "weight",
                 "q13_1"="q2", "q13_1net"="q2net"))

#joinlist2 includes all variables except for pweight/weight
joinlist1992 <- c("WP3"="AW4", "ID", "WEEK",
                  'Z7', 'Z8', 'Z9',
                  'Z10','Z10A', 'Z11',
                  'Z11A', 'SEX', 'METRO',
                  'REGION', 'HWEIGHT','STATE')

d1992 <- full_join(d35prep,
                   d34prep,
                   by=c(joinlist1992, "PWEIGHT"="WEIGHT")) %>%
  full_join(d33prep,
            by=c(joinlist1992,"PWEIGHT")) %>%
  full_join(d32prep,
            by=c(joinlist1992, "PWEIGHT"="WEIGHT")) %>%
  full_join(d31prep,
            by=c(joinlist1992, "PWEIGHT")) %>%
  full_join(d30prep,
            by=c(joinlist1992, "PWEIGHT"))

# section 4: recoding vars pre-cross year merge ----------------------------------------------
#recoding and merging 2020 variables to conform to 1992 var cases
#converts 1992 presidential vote into bush, clinton, other, or na, turning other into na later
```

```
d1992$WP3 <- recode(d1992$WP3,
                "1"="Bush", "2"="Clinton",
                "3"="Other/NA", "0"="Other/NA",
                "7"="Other/NA","8"="Other/NA",
                "9"="Other/NA",.default="Other/NA")


d2020$q13_1net <- recode(d2020$q13_1net,
                    "Biden"="Biden", "Biden and Harris"="Biden",
                    "Trump"="Trump", "Trump and Pence"="Trump",
                    "DK/No Opinion"="Other/NA", "Neither"="Other/NA",
                    "Other Candidate"="Other/NA", "Would not vote"="Other/NA")

d1992<- d1992 %>% mutate(racenet = case_when(
  (d1992$Z10==1) ~ "Hispanic",
  (d1992$Z10!=1 & d1992$Z11==1) ~ "white",
  (d1992$Z10!=1 & d1992$Z11==2) ~ "black",
  (d1992$Z10!=1 & d1992$Z11==5) ~ "OtherNA",
  (d1992$Z10!=1 & d1992$Z11==9) ~ "OtherNA"))

d1992$Z11A <- recode(d1992$Z11A,
                 '1'="a Republican", '2'="a Democrat",
                 '3'="an independent", '0'="Other/NA",
                 '8'="Other/NA",'9'="Other/NA")

d1992$SEX <- recode(d1992$SEX,
                "1"="Male", "2"="Female")

d1992$REGION <- recode(d1992$REGION,
                   "1"="Northeast", "2"="Midwest",
                   "3"="South","4"="West")

d1992$Z9 <- recode(d1992$Z9,
               "1"="Under 10 thousand dollars",
               "2"="10 to under 15 thousand dollars",
               "3"="15 to under 20 thousand dollars",
               "4"="20 to under 25 thousand dollars",
               "5"="25 to under 30 thousand dollars",
               "6"="30 to under 40 thousand dollars",
               "7"="40 to under 50 thousand dollars",
               "8"="50 to under 75 thousand dollars",
               "9"="75 thousand or more",
               "98"="Other/NA","99"="Other/NA")


d1992$METRO <- recode(d1992$METRO, "1"="M", "2"="NM")
```

```
d1992$Z8 <- recode(d1992$Z8,
               "1"="Some high school", "2"="Graduated high school",
               "3"="Some college", "4"="Graduated College",
               "5"="Post graduate", "6"="Technical school/other",
               "9"="Other/NA")

d2020$nusr <- recode(d2020$nusr,
                "R"="NM", "S"="M",
                "U"="M")


#also adding a variable to describe origin year of each dataset
d1992 <- d1992 %>% mutate(year = 1992)
d2020 <- d2020 %>% mutate(year = 2020)
```

# section 5: merging 1992 and 2020 data, and dropping unused variables-----------------------

```
joinlist3 <- c("q13_1net"="WP3", "q910"="Z7",
            "income"="Z9", "racenet",
            "partlean"="Z11A", "q924net"="SEX",
            "nusr"="METRO", "nreg4"="REGION",
            "q909"="Z8", "weight"="PWEIGHT", "year")

#dropping variables i'm not using
d2020 <- d2020 %>% select(!(respo | project | date8 |
                        q13_1 | q924 | income2 |
                        reg4 | censdiv | usr |
                        q918 | hisprace | q909a |
                        edubreak | colleduc | educnew |
                        abcnum | stcode | stcode2 |
                        pidwgt | ncensdiv))

d1992<-d1992 %>% select(!(ID | WEEK | STATE |
                      Z10 | Z10A | Z11 |
                      HWEIGHT))

#merging 1992 and 2020 datasets
fulldata <- full_join(d2020, d1992, joinlist3)
```

# section 6: renaming columns for readability --

```
colnames(fulldata)[1] <- "partyID"
colnames(fulldata)[2] <- "presVote"
colnames(fulldata)[3] <- "age"
colnames(fulldata)[4] <- "education"
#column 5 is already named income
```

```
colnames(fulldata)[6] <- "race"
colnames(fulldata)[7] <- "sex"
#column 8 is population weighting
colnames(fulldata)[9] <- "region"
colnames(fulldata)[10] <- "metropolitan"
#column 11 is year

# section 7: recoding certain variables to fix duplicates postmerge ------------------
fulldata$race <- recode(fulldata$race,
                        "Asian"="asian",
                        "Hispanic"="hispanic",
                        "other"=NA_character_,
                        "OtherNA"=NA_character_,
                        "DK/No opinion"=NA_character_,
                        " "=NA_character_)

fulldata$presVote <- recode(fulldata$presVote,
                            " "=NA_character_,
                            "Other/NA"=NA_character_)

fulldata$education <- recode(fulldata$education,
                             "(VOL) DK/No opinion"=NA_character_,
                             "(VOL) NA/Refused"=NA_character_,
                             "Some college (ASK IF TECHNICAL SCHOOL; IF YES, PUNCH CODE 3,
FOR HIGH SCHOOL)"="Some college",
                             "Other/NA"=NA_character_,
                             "8th grade or less"="8th grade or less",
                             "Graduated College"="Graduated College",
                             "Graduated high school"="Graduated high school",
                             "Post graduate"="Post graduate",
                             "Some college"="Some college",
                             "Some high school"="Some high school",
                             "Technical school/other"="Technical school/other",
                             .default=NA_character_)

fulldata <- fulldata %>% mutate(metropolitan = case_when(
  metropolitan=="M" ~ "M",
  metropolitan=="NM" ~ "NM",
  TRUE ~ NA_character_
))
fulldata$race <- relevel(factor(fulldata$race), ref="white") #making it so that white,
nonmetropolitan is the intercept case
fulldata$metropolitan <- relevel(factor(fulldata$metropolitan), ref="NM")


# section 8: logit  ------------------------------------------------------------------
```

```
md1992 <- fulldata %>%
  filter(year=="1992") %>%
  na.omit(cols=c("presVote","race","metropolitan"))

md2020 <- fulldata %>%
  filter(year=="2020") %>%
  na.omit(cols=c("presVote","race","metropolitan"))


#recoding presvote for 1992 so clinton/dem=success(1), and bush/repub=failure(0)

md1992$presVote <- recode(md1992$presVote,
                          "Clinton"=1, "Bush"=0)
#recoding presvote for 1992 so biden/dem=success(1), and trump/repub=failure(0)

md2020$presVote <- recode(md2020$presVote,
                          "Biden"=1, "Trump"=0)
#presvote is my response var, with metro and race as my explanatory vars

m1992prm <- glm(presVote ~ factor(race) + factor(metropolitan),
                family = binomial(link = 'logit'),
                data = md1992)

m2020prm <- glm(presVote ~ factor(race) + factor(metropolitan),
                family = binomial(link = 'logit'),
                data = md2020)

summary(m1992prm)
summary(m2020prm)
exp(coef(m1992prm))[-1] #gives odds ratios for 1992
exp(coef(m2020prm))[-1] #odds ratios 2020

#confirms that for both 1992 and 2020, biden and clinton both received the majority of the Black
vote in my dataset, as was in reality
table(md1992$presVote, md1992$race)
table(md2020$presVote, md2020$race)
table(md1992$presVote, md1992$metro)
table(md2020$presVote, md2020$metro)
```

Reflection

Over the course of doing this project I learned many important skills. One part of the work I did consisted of improving my technical skills. I learned how to process, clean, and translate legacy data, taking longitudinal legacy data and modifying it to run with a modern dataset. I gained experience with dplyr and learned a new package called rio to convert my files. The other side of the skills I learned was significantly less technical. Originally, I had proposed using all of the polling data available, for every election possible, for my analysis. This was very foolish. That project would have been significantly beyond the scope of this class, and more likely in the realm of a Ph.D. or Master’s thesis. As it was, cleaning and processing the datasets and joining them all together took me nearly 2 months for just two years worth of data. I got much better at self editing during this project. I can have a great idea, but I need to be able to whittle it down into usable, compleatable, and manageable parts. Additionally, this project has helped me improve at crafting answerable questions. Going forward, I will be more discerning in how I select data and craft projects, especially with regard to collecting data, and even more so with legacy data.